Text-informed audio source separation. Example-based approach using non-negative matrix partial co-factorization

Luc Le Magoarou, Alexey Ozerov, Ngoc Duong

Noname manuscript No.
(will be inserted by the editor)

Text-informed audio source separation. Example-based approach using non-negative matrix partial co-factorization

Luc Le Magoarou · Alexey Ozerov · Ngoc Q. K. Duong

Received: date / Accepted: date

Abstract The so-called *informed* audio source separation, where the separation process is guided by some auxiliary information, has recently attracted a lot of research interest since classical *blind* or *non-informed* approaches often do not lead to satisfactory performances in many practical applications. In this paper we present a novel *text-informed* framework in which a target speech source can be separated from the background in the mixture using the corresponding textual information. First, given the text, we propose to produce a speech example via either a speech synthesizer or a human. We then use this example to guide source separation and, for that purpose, we introduce a new variant of the non-negative matrix partial co-factorization (NMPCF) model based on a so-called *excitation-filter-channel* speech model. Such a modeling allows sharing the linguistic information between the speech example and the speech in the mixture. The corresponding multiplicative update (MU) rules are eventually derived for the parameters estimation and several extensions of the model are proposed and investigated. We perform extensive experiments to assess the effectiveness of the proposed approach in terms of source separation and alignment performance.

Luc Le Magoarou
Inria Rennes - Bretagne Atlantique, Campus universitaire de Beaulieu 35042 Rennes Cedex, France
E-mail: luc.le-magoarou@inria.fr

Alexey Ozerov and Ngoc Q. K. Duong
Technicolor, 975 avenue des Champs Blancs, CS 17616, 35576 Cesson Sévigné, France E-mail: {alexey.ozerov, quang-khanh-ngoc.duong}@technicolor.com

Most of this work was done while the first author was with Technicolor, and a part of the work has been presented at the 2013 IEEE International Workshop on Machine Learning for Signal Processing (MLSP) [1].

Keywords Text-informed audio source separation, Non-negative Matrix Partial Co-Factorization, excitation-filter model, speech alignment

1 Introduction

Audio source separation, which aims at extracting individual sound sources from the observed mixture signal, offers a wide range of applications in, *e.g.*, automatic speech recognition, hearing aids, movie dubbing, and so on. However, despite a lot of research effort, *blind* source separation still does not provide a satisfactory performance, and is difficult especially in under-determined cases where the number of sources exceeds the number of observed mixtures [2]. An emerging research trend, referred to as *informed* source separation, has been widely considered recently and was shown to be highly effective for certain source separation tasks. It consists in using some auxiliary information about the sources and/or the mixing process to guide the separation. For example, *score-informed* approaches rely on musical score to guide the separation in music recordings [3–6], separation-by-humming (SbH) algorithms exploit a sound "hummed" by the user mimicking the source of interest [7,8], and user-guided approaches take into account knowledge about, *e.g.*, user-selected F0 track [9] or user-annotated source activity patterns along the spectrogram of the mixture [10,11] and/or that of the estimated sources [12,13]. In line with this direction, there are also speech separation systems informed, *e.g.*, by speaker gender [14], by corresponding video [15], or by the natural language structure [16]. However, while written text corresponding to the speech in the mixture is often available, *e.g.*, in form of subtitles (an approximate speech transcrip-

tion) associated to a movie or script (an exact speech transcription) in a movie production phase, to the best of our knowledge, none of the existing approaches, except our preliminary work [1], exploits this information to guide the separation process.

With the above mentioned motivation, we introduce in this paper a novel framework that exploits the available textual information to guide the separation of target speech from the background in single-channel mixtures. The proposed approach is inspired by the synthesis-based score-informed music separation approaches [3,6] where a symbolic representation of the corresponding music sources (the score) is used to synthesize audio examples that are further used to guide the separation. In our scenario, the available text is used to generate a speech example, *e.g.*, via a speech synthesizer, which shares the linguistic information with the speech in the mixture (since the example and the speech in the mixture contain the same uttered text). Note that, as compared to the music case in [3,6], such a task is intrinsically more challenging for speech. Indeed, in contrast to music, where the temporal mismatch between the sources and the score-synthesized examples is usually linear (the tempo may not be the same, but the rhythm is usually maintained), it is often non-linear for speech. Moreover, while the pitches of the same musical notes are usually on the same frequency locations, there is no guarantee that the pitches of two different speakers would be the same. In order to handle such kind of variations in both frequency and time between the latent source and the synthesized speech example, we develop a novel variant of the non-negative matrix partial co-factorization (NMPCF) model [1]. The proposed model is based on a so-called *excitation-filter-channel* (EFC), which is a new extension of the excitation-filter model [18,19]. This formulation allows to jointly factorize the spectrogram of the speech example and that of the mixture, while sharing between them the common linguistic information and handling the variations of both the temporal dynamics, the recording conditions, and the speaker's prosody and timber.

As compared to our preliminary work [1], this paper contains the following main additional contributions:

- On the methodological level, new structural constraints (inspired by [20]) on some matrices of the NMPCF model are introduced and investigated. These constraints have quite natural physical motivations. The first type of constraint, imposed on some matrix of activation coefficients, means that at most one element from the corresponding dictionary can be active at a given time, which can be viewed as a sort of extreme sparsity. A physical motivation behind is that in a monophonic speech signal at most one phoneme can be pronounced at a time and at most one pitch can be active at a time. The second constraint is imposed on a so-called synchronization matrix and simply means that the synchronization between the example and the speech in the mixture must be monotonous.
- On the experimental level, we study in depth the influence of some key parameters on the separation performance and assess explicitly the benefit of the introduced excitation-filter-channel model compared to the state-of-the-art excitation-filter model. Furthermore, we evaluate the potential of the considered model in speech alignment task.

Moreover, we present in this manuscript a full list of equations for parameter updates. Finally, note that the proposed framework is applicable in both single-channel and multichannel mixtures. However, for simplicity we focus the presentation on the single-channel case in this paper. Extending the approach to the multichannel case is quite straightforward (*e.g.*, one can be inspired by developments in [20] to do so). We have done it in practice by combining the considered spectral model with the spatial model introduced in [21] in order to participate in the "Two-channel mixtures of speech and real-world background noise" separation task of the Signal Separation Evaluation Campaign (SiSEC 2013) [22].

The structure of the rest of the paper is as follows. A general workflow of the proposed framework and some related work are presented in section 2. The NMPCF-based modeling as well as new structural constraints are then described in Section 3, followed by presentation of the model parameter estimation algorithms in section 4. The proposed approach is extensively studied and evaluated in various settings in terms of both source separation and speech alignment performances in section 5. Finally we conclude in section 6.

2 General workflow and related work

The general workflow of the proposed approach is depicted in Fig. 1 where a speech example corresponding to the same uttered words as the one of the speech in the mixture is assumed to be available. The proposed source separation algorithm takes as input the observed audio mixture and the speech example to guide the separation. The NMPCF model and the corresponding parameter estimation block will be described in details in sections 3 and 4, respectively. Finally, the targeted speech and background estimates are recon-

[1] NMPCF model [5] is a particular case of a more general generalized coupled tensor factorization (GCTF) model that was used as well for informed source separation [17].

structed from the estimated parameters via the standard Wiener filtering [23] as described at the end of section 3.

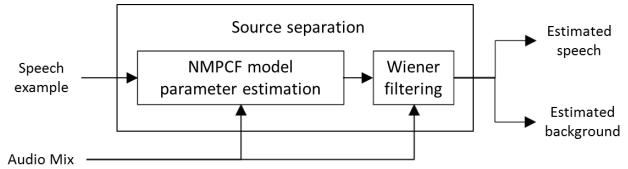

Fig. 1 General workflow of the proposed approach.

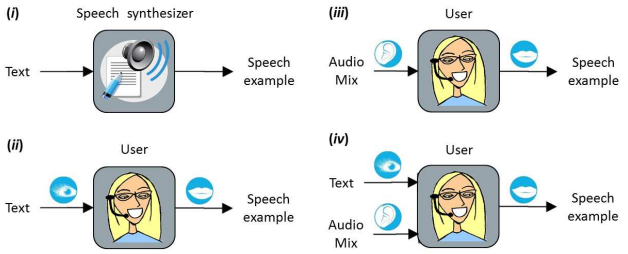

Fig. 2 Possible ways of speech example production.

One can imagine several ways to generate a speech example that carries similar linguistic information to the speech in the mixture. We identify in the following four strategies to produce such an example (see Fig. 2). The first one (*i*) uses the text often provided with TV programs and DVD movies (subtitles or script) to produce a speech example using an automatic speech synthesizer. This scenario is probably among the easiest ones since it is totally automatic and does not require any intervention from the user. The three other ways we consider are semi-automatic and need the user speaking to produce the example. Depending on the availability of the information and on user's wishes, he/she can either (*ii*) simply read the text, (*iii*) mimic the speech in the mixture after having listened to it, or (*iv*) do both. In summary, we see that besides introducing the text-informed approach we introduce in fact several practical methods lying in-between the text-informed and the user-guided approaches. As such, method (*i*) is purely text-informed and method (*iii*) is purely user-guided, while methods (*ii*) and (*iv*) are in-between, since rely on both the text availability and an intervention from user.

Though the considered speech example-based text-informed strategies have not been presented yet in the existing works, except our preliminary study [1], some related approaches would be worth to be mentioned. Pedone *et al.* [24] proposed an algorithm for phoneme-level text to audio synchronisation applied to mixtures of speech and background music. This algorithm relies on the non-negative matrix factorization (NMF)-like framework where the excitation-filter models of English phonemes are pre-trained. In this light, given that there is a sort of latent speech modeling guided by text, it could be extended as well for text-informed speech separation. However, while this approach was not evaluated in terms of source separation and requires to learn the general phoneme models, our method exploits specific phonemes in a speech example, which is probably pronounced in a closer way to the speech in the mixture. By this difference, we believe that the proposed approach potentially brings better separation performance. Using a sound mimicking the one to be extracted from the mixture to guide the separation, Smaragdis *et al.* introduced a so-called *Separation by Humming (SbH)* approach based on the probabilistic latent component analysis (PLCA) [7] and FitzGerald [8] reported a similar method based on the NMF. However, while the performance resulted from PLCA or NMF [8] is limited due to the strong variations between the source and the example (*e.g.*, pitch or temporal dynamic variation, as mentioned in the introduction), our proposed NMPCF framework models those variations explicitly.

3 Modeling framework

We first formulate the mixing problem and we describe separately the proposed excitation-filter-channel spectral models of the mixture and the speech example as well as explain why we chose this specific models. Finally we the present the NMPCF-based [5] couplings between these two models. We further introduce novel structural constraints into the NMPCF model. Finally, we explain how the sources can be reconstructed given the estimated model parameters.

3.1 Problem formulation

Let us consider a single-channel mixture:

$$x(t) = s(t) + b(t) \tag{1}$$

consisting of a target speech signal $s(t)$ corrupted by a background signal $b(t)$, where t denotes the discrete time index. The goal is to estimate speech, given the mixture $x(t)$ and a speech example $y(t)$.

3.2 Mixture model

Let $\mathbf{X} \in \mathbb{C}^{F \times N}$ be the Short-Time Fourier Transform (STFT) of $x(t)$, F being the number of frequency bins

and N the number of time frames. Equation (1) rewrites:

$$\mathbf{X} = \mathbf{S} + \mathbf{B}, \tag{2}$$

where $\mathbf{S}$ and $\mathbf{B}$ are the STFTs of the speech and the background, respectively. Defining the power spectrogram $\mathbf{V}_X = |\mathbf{X}|^{\cdot[2]}$ ($\mathbf{A}^{\cdot[b]}$ being the element-wise exponentiation of a matrix $\mathbf{A}$ by b), assuming that the speech and background signals are uncorrelated, $\mathbf{V}_X$ can be approximated as:

$$\mathbf{V}_X \approx \hat{\mathbf{V}}_X = \hat{\mathbf{V}}_S + \hat{\mathbf{V}}_B, \tag{3}$$

where $\hat{\mathbf{V}}_X, \hat{\mathbf{V}}_S, \hat{\mathbf{V}}_B \in \mathbb{R}_+^{F\times N}$ are approximations of the power spectrograms of the mixture, the speech and the background, respectively.

We further constrain the speech by imposing a so-called *excitation-filter-channel (EFC)* [2] structure on $\hat{\mathbf{V}}_S$:

$$\hat{\mathbf{V}}_S = \hat{\mathbf{V}}_S^e \odot \hat{\mathbf{V}}_S^\phi \odot \hat{\mathbf{V}}_S^c, \tag{4}$$

with $\odot$ being the Hadamard element-wise product, $\hat{\mathbf{V}}_S^e$ being a time-varying linear combination of comb filters modeling the pitch, $\hat{\mathbf{V}}_S^\phi$ being a time-varying filter modeling the phonemes pronounced, and $\hat{\mathbf{V}}_S^c$ being a time-invariant filter modeling the recording conditions and speaker's vocal tract. Let us stress that, except if the contrary is stated, all the entries of matrices within power spectrogram models in this paper are assumed real and non-negative numbers.

All the matrices in Eq. (4) and matrix $\hat{\mathbf{V}}_B$ are further subject to NMF decompositions as follows:

- $\hat{\mathbf{V}}_S^e = \mathbf{W}^e\mathbf{H}_S^e$, $\mathbf{W}^e \in \mathbb{R}_+^{F\times I}$ being a pre-defined dictionary of combs representing all possible pitches of human voice and $\mathbf{H}_S^e \in \mathbb{R}_+^{I\times N}$ being the corresponding temporal activations.
- $\hat{\mathbf{V}}_S^\phi = \mathbf{W}_S^\phi\mathbf{H}_S^\phi$, $\mathbf{W}_S^\phi \in \mathbb{R}_+^{F\times J}$ being a dictionary of phoneme spectral envelopes and $\mathbf{H}_S^\phi \in \mathbb{R}_+^{J\times N}$ being the corresponding temporal activations.
- $\hat{\mathbf{V}}_S^c = \mathbf{w}_S^c\mathbf{i}_N^T$, $\mathbf{w}_S^c \in \mathbb{R}_+^{F\times 1}$ modeling both the spectral shape of the recording conditions filter and speaker's vocal tract, and $\mathbf{i}_N$ being an N-length column vector of ones.
- $\hat{\mathbf{V}}_B = \mathbf{W}_B\mathbf{H}_B$, $\mathbf{W}_B \in \mathbb{R}_+^{F\times K}$ being a dictionary of background spectral shapes and $\mathbf{H}_B \in \mathbb{R}_+^{K\times N}$ being the corresponding temporal activations.

Another assumption is made so as to constrain spectral shapes of matrices $\mathbf{W}_S^\phi$ and $\mathbf{w}_S^c$ to be smooth [20]. Following [20], these matrices are constrained as follows : $\mathbf{W}_S^\phi = \mathbf{P}\mathbf{E}_S^\phi$ and $\mathbf{w}_S^c = \mathbf{P}\mathbf{e}_S^c$, where $\mathbf{P} \in \mathbb{R}_+^{F\times L}$ is a pre-defined matrix of L so-called *spectral blobs*, that are used to construct $\mathbf{W}_S^\phi$ and $\mathbf{w}_S^c$ with weights $\mathbf{E}_S^\phi \in \mathbb{R}_+^{L\times J}$ and $\mathbf{e}_S^c \in \mathbb{R}_+^{L\times 1}$, respectively.

Finally, the mixture model can be summarized as:

$$\mathbf{V}_X \approx \hat{\mathbf{V}}_X = \overbrace{\underbrace{\left(\mathbf{W}^e\mathbf{H}_S^e\right)}_{\hat{\mathbf{V}}_S^e} \odot \underbrace{\left(\mathbf{W}_S^\phi\mathbf{H}_S^\phi\right)}_{\hat{\mathbf{V}}_S^\phi} \odot \underbrace{\left(\mathbf{w}_S^c\mathbf{i}_N^T\right)}_{\hat{\mathbf{V}}_S^c}}^{\hat{\mathbf{V}}_S} + \overbrace{\mathbf{W}_B\mathbf{H}_B}^{\hat{\mathbf{V}}_B}. \tag{5}$$

3.3 Speech example model

Let $\mathbf{Y} \in \mathbb{C}^{F\times N'}$ be the STFT of $y(t)$ and $\mathbf{V}_Y = |\mathbf{Y}|^{\cdot[2]} \in \mathbb{R}_+^{F\times N'}$ its power spectrogram. Note that in most cases $N' \neq N$ due to the temporal mismatch between the example and the mixture. The example consists of only one clean speech source whose power spectrogram is approximated as:

$$\mathbf{V}_Y \approx \hat{\mathbf{V}}_Y = \hat{\mathbf{V}}_Y^e \odot \hat{\mathbf{V}}_Y^\phi \odot \hat{\mathbf{V}}_Y^c, \tag{6}$$

where $\hat{\mathbf{V}}_Y^e$, $\hat{\mathbf{V}}_Y^\phi$ and $\hat{\mathbf{V}}_Y^c$ are decomposed the same way as in section 3.2, i.e., $\hat{\mathbf{V}}_Y^e = \mathbf{W}^e\mathbf{H}_Y^e$, $\hat{\mathbf{V}}_Y^\phi = \mathbf{W}_Y^\phi\mathbf{H}_Y^\phi$ and $\hat{\mathbf{V}}_Y^c = \mathbf{w}_Y^c\mathbf{i}_{N'}^T$. The smoothness constraints are applied as well: $\mathbf{W}_Y^\phi = \mathbf{P}\mathbf{E}_Y^\phi$ and $\mathbf{w}_Y^c = \mathbf{P}\mathbf{e}_Y^c$.

As a result, the spectrogram of the speech example is modeled as:

$$\mathbf{V}_Y \approx \hat{\mathbf{V}}_Y = \underbrace{\left(\mathbf{W}^e\mathbf{H}_Y^e\right)}_{\hat{\mathbf{V}}_Y^e} \odot \underbrace{\left(\mathbf{W}_Y^\phi\mathbf{H}_Y^\phi\right)}_{\hat{\mathbf{V}}_Y^\phi} \odot \underbrace{\left(\mathbf{w}_Y^c\mathbf{i}_{N'}^T\right)}_{\hat{\mathbf{V}}_Y^c}. \tag{7}$$

3.4 Couplings between the mixture and example models

The role of the example is to guide source separation, thanks to the fact that it shares common linguistic information with the speech in the mixture. We model this sharing as follows:

- The phonemes pronounced in the mixture and those pronounced in the example are the same, thus we assume: $\mathbf{W}_S^\phi = \mathbf{W}_Y^\phi = \mathbf{W}^\phi$, and $\mathbf{W}^\phi$ is to be estimated. This assumption implies $\mathbf{E}_S^\phi = \mathbf{E}_Y^\phi = \mathbf{E}^\phi$.
- The phonemes are pronounced in the same order in the mix and in the example, but not exactly temporally synchronized. Thus we represent $\mathbf{H}_S^\phi$ as $\mathbf{H}_S^\phi = \mathbf{H}_Y^\phi\mathbf{D}$ where $\mathbf{D} \in \mathbb{R}_+^{N'\times N}$ is a so-called *synchronization matrix* [24]. $\mathbf{H}_Y^\phi$ and $\mathbf{D}$ are to be estimated.

The synchronization matrix D is constrained to be non-zero only within a vertical band of size B around an initial synchronization path found by a Dynamic Time

[2] The proposed EFC model is a new extension of the excitation-filter model [19].

Warping (DTW) algorithm [25] applied on the Mel-Frequency Cepstral Coefficients (MFCC) representations of the example and the mixture (see Fig. 4, bottom-left subplot). This constraint means that only a small desynchronization (whose maximal amplitude is specified by B) from the initial DTW path is allowed. Thanks to the property of the multiplicative update rules [23] we use for parameter estimation (see section 4.2 below) to keep zero parameters unchanged, to maintain such a constraint it is enough to simply initialize matrix D as such. We have found experimentally in [1] that this strategy was the best among other tested strategies, and the influence of the bandwidth B on the separation performance is studied experimentally in section 5.3.1 below.

It is worth noting that the above assumptions are reasonable since the mixture and the example contain utterances of the same sentences. The final NMPCF model is as follows:

$$\begin{aligned}\mathbf{V}_Y \approx \hat{\mathbf{V}}_Y &= \left(\mathbf{W}^e\mathbf{H}^e_Y\right) \odot \left(\mathbf{W}^\phi\mathbf{H}^\phi_Y\right) \quad \odot\left(\mathbf{w}^c_Y\mathbf{i}^T_{N'}\right),\\ \mathbf{V}_X \approx \hat{\mathbf{V}}_X &= \left(\mathbf{W}^e\mathbf{H}^e_S\right) \odot \left(\mathbf{W}^\phi\mathbf{H}^\phi_Y\mathbf{D}\right)\odot\left(\mathbf{w}^c_S\mathbf{i}^T_N\right) + \mathbf{W}_B\mathbf{H}_B,\end{aligned} \tag{8}$$

where pre-defined and fixed parameters are $\mathbf{W}^e$, $\mathbf{i}^T_{N'}$ and $\mathbf{i}^T_N$ (in green), shared and estimated parameters are $\mathbf{W}^\phi$ and $\mathbf{H}^\phi_Y$ (in red), and non-shared and estimated parameters are the others (in black). These couplings are visualized on Fig. 3.

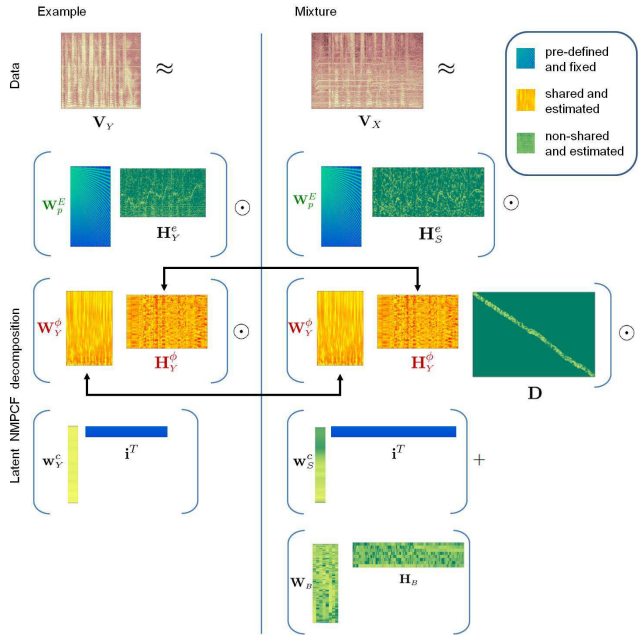

Fig. 3 NMPCF model for the speech example and the mixture.

To summarize, the parameters to be estimated are [3]:

$$\boldsymbol{\theta} = \left\{\mathbf{H}^e_Y, \mathbf{H}^e_S, \mathbf{E}^\phi, \mathbf{H}^\phi_Y, \mathbf{D}, \mathbf{e}^c_Y, \mathbf{e}^c_S, \mathbf{H}_B, \mathbf{W}_B\right\}, \tag{9}$$

while $\mathbf{W}^e$, $\mathbf{P}$, $\mathbf{i}^T_N$ and $\mathbf{i}^T_{N'}$ are pre-defined and fixed.

3.5 Structural constraints on the model parameters

We consider further constraints on the matrices constituting the spectral model in order to better fit the physical phenomena. It would for example make sense to constrain the speech to have only one active fundamental frequency at time, and also only one active phoneme at each time. These types of constraints can be translated within our modeling by allowing only one non-zero entry in each column of the matrices $\mathbf{H}^e_Y$, $\mathbf{H}^e_S$ and $\mathbf{H}^\phi_Y$, respectively (see Fig. 4, top-right and middle-right subplots). Such constraints have been considered in [20] under the name of Gaussian Scaled Mixture Model (GSMM).

In the same spirit, we can also further constrain the synchronization matrix $\mathbf{D}$ by allowing only one path from the top-left to the bottom-right corner to be non-zero (see Fig. 4, bottom-right subplot). This constraint means that we still allow a desynchronization of the path in $\mathbf{D}$ from the initial DTW path, but we require it to be monotonous as a DTW. Let us stress however that this constraint is not equivalent to setting the bandwidth $B = 1$, the desired non-zero path in $\mathbf{D}$ being still allowed to vary within a bandwidth $B \geq 1$ around the initial DTW path. We call this constraint "D-Struct" below.

We study the influence of all these constraints on the separation performance and show experimental results in section 5.3.3 below.

3.6 Source reconstruction via Wiener filtering

Given the NMPCF model parameters $\boldsymbol{\theta}$ (Eq. (9)) estimated, as described in the following section, a speech source STFT estimate $\hat{\mathbf{S}} \in \mathbb{C}^{F\times N}$ is computed by the standard Wiener filtering as:

$$\hat{\mathbf{S}} = \left(\hat{\mathbf{V}}_S \,/\, \hat{\mathbf{V}}_X\right) \odot \mathbf{X}, \tag{10}$$

where all the operations are element-wise, $\mathbf{X}$ is the mixture STFT; and $\hat{\mathbf{V}}_S$ and $\hat{\mathbf{V}}_X$ are computed, respectively, as in (3) and (4), given $\boldsymbol{\theta}$. A background source estimate is simply computed as $\hat{\mathbf{B}} = \mathbf{X} - \hat{\mathbf{S}}$, and the corresponding time signals may then be obtained through inverse STFT using an adequate overlap-add procedure with dual synthesis window.

[3] Keep in mind that $\mathbf{W}^\phi = \mathbf{P}\mathbf{E}^\phi$, $\mathbf{w}^c_S = \mathbf{P}\mathbf{e}^c_S$ and $\mathbf{w}^c_Y = \mathbf{P}\mathbf{e}^c_Y$.

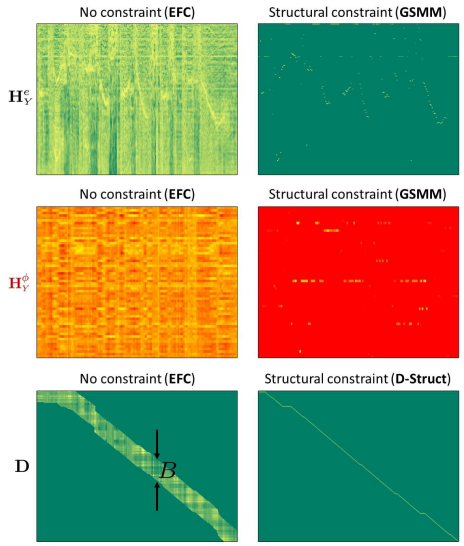

Fig. 4 Illustration of new structural constraints on matrices $\mathbf{H}_Y^e$ (only one pitch can be active at time), $\mathbf{H}_Y^\phi$ (only one phoneme can be pronounced at time and $\mathbf{D}$ (the alignment between the example and the speech in the mixture must be monotonous). Matrices without (as those in Fig. 3) and with structural constraints are represented in the left and right columns, respectively.

4 Parameter estimation

This section is devoted to the estimation of model parameters $\boldsymbol{\theta}$ summarized in Eq. (9). We first introduce the optimization cost in section 4.1. We then provide the corresponding Multiplicative Update (MU) rules in section 4.2 and describe some hints so as to how the structural constraints presented in Section 3.5 can be taken into account during the estimation process.

4.1 Cost function

The general principle of NMF-like parameter estimation is to minimize certain cost function measuring a divergence between the data matrix and its structural approximation. We consider here the Itakura-Saito (IS) divergence [4] and specify the cost function as follows:

$$C(\boldsymbol{\theta}) = \sum_{f,n=1}^{F,N} d_{IS}\left(v_{X,fn}|\hat{v}_{X,fn}\right) + \lambda \sum_{f,n=1}^{F,N'} d_{IS}\left(v_{Y,fn}|\hat{v}_{Y,fn}\right), \tag{11}$$

where $\lambda \in \mathbb{R}_+$ is a trade-off parameter that determines the example's influence on the estimation, $d_{IS}\left(a|b\right) = a/b - \log\left(a/b\right) - 1$ is the IS divergence, $v_{Y,fn}$, $v_{X,fn}$, $\hat{v}_{Y,fn}$ and $\hat{v}_{X,fn}$ are, respectively, entries of data matrices $\mathbf{V}_Y$, $\mathbf{V}_X$ and their structural approximations $\hat{\mathbf{V}}_Y$, $\hat{\mathbf{V}}_X$ from (8).

[4] When applied to power spectrograms of audio signals, IS divergence was shown as one of the most suitable choices for NMF-like decompositions [23], in particular thanks to its scale invariance property.

4.2 Parameter estimation via MU rules

To optimize the cost (11) we use standard multiplicative update (MU) rules which can be derived following a recipe described in [23]. The idea is to derive MU rules based on the cost function's gradient with respect to each parameter. Most of the resulting updates are very similar to those described, e.g., in [20], and are as follows:

$$\mathbf{H}_Y^e \leftarrow \mathbf{H}_Y^e \odot \frac{\mathbf{W}^{eT}\left[\left(\mathbf{W}_Y^\phi \mathbf{H}_Y^\phi\right) \odot \left(\mathbf{w}_Y^c \mathbf{i}_{N'}^T\right) \odot \hat{\mathbf{V}}_Y^{[-2]} \odot \mathbf{V}_Y\right]}{\mathbf{W}^{eT}\left[\left(\mathbf{W}_Y^\phi \mathbf{H}_Y^\phi\right) \odot \left(\mathbf{w}_Y^c \mathbf{i}_{N'}^T\right) \odot \hat{\mathbf{V}}_Y^{[-1]}\right]} \tag{12}$$

$$\mathbf{H}_S^e \leftarrow \mathbf{H}_S^e \odot \frac{\mathbf{W}^{eT}\left[\left(\mathbf{W}_S^\phi \mathbf{H}_S^\phi\right) \odot \left(\mathbf{w}_S^c \mathbf{i}_N^T\right) \odot \hat{\mathbf{V}}_X^{[-2]} \odot \mathbf{V}_X\right]}{\mathbf{W}^{eT}\left[\left(\mathbf{W}_S^\phi \mathbf{H}_S^\phi\right) \odot \left(\mathbf{w}_S^c \mathbf{i}_N^T\right) \odot \hat{\mathbf{V}}_X^{[-1]}\right]} \tag{13}$$

$$\mathbf{E}^\phi \leftarrow \mathbf{E}^\phi \odot \frac{\mathbf{P}^T\left[\lambda\left(\left(\mathbf{W}^e\mathbf{H}_Y^e\right) \odot \left(\mathbf{w}_Y^c \mathbf{i}_{N'}^T\right) \odot \hat{\mathbf{V}}_Y^{[-2]} \odot \mathbf{V}_Y\right)\mathbf{H}_Y^{\phi\,T} + \left(\left(\mathbf{W}^e\mathbf{H}_S^e\right) \odot \left(\mathbf{w}_S^c \mathbf{i}_N^T\right) \odot \hat{\mathbf{V}}_X^{[-2]} \odot \mathbf{V}_X\right)\mathbf{H}_S^{\phi\,T}\right]}{\mathbf{P}^T\left[\lambda\left(\left(\mathbf{W}^e\mathbf{H}_Y^e\right) \odot \left(\mathbf{w}_Y^c \mathbf{i}_{N'}^T\right) \odot \hat{\mathbf{V}}_Y^{[-1]}\right)\mathbf{H}_Y^{\phi\,T} + \left(\left(\mathbf{W}^e\mathbf{H}_S^e\right) \odot \left(\mathbf{w}_S^c \mathbf{i}_N^T\right) \odot \hat{\mathbf{V}}_X^{[-1]}\right)\mathbf{H}_S^{\phi\,T}\right]} \tag{14}$$

$$\mathbf{H}_Y^\phi \leftarrow \mathbf{H}_Y^\phi \odot \frac{\lambda\mathbf{W}_Y^{\phi\,T}\left(\left(\mathbf{W}^e\mathbf{H}_Y^e\right) \odot \left(\mathbf{w}_Y^c \mathbf{i}_{N'}^T\right) \odot \hat{\mathbf{V}}_Y^{[-2]} \odot \mathbf{V}_Y\right) + \mathbf{W}_S^{\phi\,T}\left(\left(\mathbf{W}^e\mathbf{H}_S^e\right) \odot \left(\mathbf{w}_S^c \mathbf{i}_N^T\right) \odot \hat{\mathbf{V}}_X^{[-2]} \odot \mathbf{V}_X\right)\mathbf{D}^T}{\lambda\mathbf{W}_Y^{\phi\,T}\left(\left(\mathbf{W}^e\mathbf{H}_Y^e\right) \odot \left(\mathbf{w}_Y^c \mathbf{i}_{N'}^T\right) \odot \hat{\mathbf{V}}_Y^{[-1]}\right) + \mathbf{W}_S^{\phi\,T}\left(\left(\mathbf{W}^e\mathbf{H}_S^e\right) \odot \left(\mathbf{w}_S^c \mathbf{i}_N^T\right) \odot \hat{\mathbf{V}}_X^{[-1]}\right)\mathbf{D}^T} \tag{15}$$

$$\mathbf{D} \leftarrow \mathbf{D} \odot \frac{\mathbf{H}_Y^{\phi\,T}\mathbf{W}_S^{\phi\,T}\left[\left(\mathbf{W}^e\mathbf{H}_S^e\right) \odot \left(\mathbf{w}_S^c \mathbf{i}_N^T\right) \odot \hat{\mathbf{V}}_X^{[-2]} \odot \mathbf{V}_X\right]}{\mathbf{H}_Y^{\phi\,T}\mathbf{W}_S^{\phi\,T}\left[\left(\mathbf{W}^e\mathbf{H}_S^e\right) \odot \left(\mathbf{w}_S^c \mathbf{i}_N^T\right) \odot \hat{\mathbf{V}}_X^{[-1]}\right]} \tag{16}$$

$$\mathbf{e}_Y^c \leftarrow \mathbf{e}_Y^c \odot \frac{\mathbf{P}^T\left[\left(\mathbf{W}^e\mathbf{H}_Y^e\right) \odot \left(\mathbf{W}_Y^\phi\mathbf{H}_Y^\phi\right) \odot \hat{\mathbf{V}}_Y^{[-2]} \odot \mathbf{V}_Y\right]\mathbf{i}_{N'}}{\mathbf{P}^T\left[\left(\mathbf{W}^e\mathbf{H}_Y^e\right) \odot \left(\mathbf{W}_Y^\phi\mathbf{H}_Y^\phi\right) \odot \hat{\mathbf{V}}_Y^{[-1]}\right]\mathbf{i}_{N'}} \tag{17}$$

$$\mathbf{e}_S^c \leftarrow \mathbf{e}_S^c \odot \frac{\mathbf{P}^T\left[\left(\mathbf{W}^e\mathbf{H}_S^e\right) \odot \left(\mathbf{W}_S^\phi\mathbf{H}_S^\phi\right) \odot \hat{\mathbf{V}}_X^{[-2]} \odot \mathbf{V}_X\right]\mathbf{i}_N}{\mathbf{P}^T\left[\left(\mathbf{W}^e\mathbf{H}_S^e\right) \odot \left(\mathbf{W}_S^\phi\mathbf{H}_S^\phi\right) \odot \hat{\mathbf{V}}_X^{[-1]}\right]\mathbf{i}_N} \tag{18}$$

$$\mathbf{H}_B \leftarrow \mathbf{H}_B \odot \frac{\mathbf{W}_B{}^T\left(\hat{\mathbf{V}}_X^{[-2]} \odot \mathbf{V}_X\right)}{\mathbf{W}_B{}^T\left(\hat{\mathbf{V}}_X^{[-1]}\right)} \tag{19}$$

$$\mathbf{W}_B \leftarrow \mathbf{W}_B \odot \frac{\left(\hat{\mathbf{V}}_X^{[-2]} \odot \mathbf{V}_X\right)\mathbf{H}_B{}^T}{\left(\hat{\mathbf{V}}_X^{[-1]}\right)\mathbf{H}_B{}^T} \tag{20}$$

Let us just note that the updates of shared parameters ($\mathbf{E}^\phi$ and $\mathbf{H}_Y^\phi$) take into account both data matrices ($\mathbf{V}_Y$ and $\mathbf{V}_X$), all other NMPCF model parameters, as well as the trade-off parameter λ.

4.3 Updates with structural constraints

Note that due to the property of MU rules to keep zero parameters unchanged, these rules are not directly applicable with the structural constraints introduced in section 3.5. Indeed, with these constraints one needs re-estimating on each iteration an appropriate support

for the corresponding matrix, while MU rules would get stuck from the first iteration in one support.

As such for updating activation matrices $\mathbf{H}_Y^e$, $\mathbf{H}_S^e$ and $\mathbf{H}_Y^\phi$, which correspond to the GSMM modeling [20], we adopt exactly the same strategy as in [20]. In summary, this strategy consists in first updating locally all the entries of one matrix using the corresponding update among (12), (13) and (15), and in choosing one entry per column yielding the highest likelihood while setting to zero all the other entries (see [20] for more details). This strategy guarantees a local optimization of the cost (11) in the sense that the cost is guaranteed to remain non-increasing after each update.

Regarding the "D-Struct" constraint for the synchronization matrix $\mathbf{D}$, we elaborated a similar process. Each entry of $\mathbf{D}$ within the admissible area, *i.e.*, within the B-bandwidth around the initial DTW path (as such we also maintain the previous constraint), is first locally updated using equation (16), and the corresponding local log-likelihood are computed. A monotonous path maximizing the corresponding accumulated log-likelihoods is then computed using the same DTW implementation [25] we used previously. However, we must acknowledge that this optimization strategy is rather ad hoc, since does not guarantee local optimization of the cost (11). We believe that making it locally-optimal would be possible, *e.g.*, by adopting a sort of Hidden Markov Model (HMM)-based decoding instead of the DTW. Nevertheless, we do not address this issue in this work and leave it for future investigation.

5 Experiments

All the experiments and / or results presented in this section are new, except the results of the baseline methods in Table 3 below that are as in our previous contribution [1]. We first describe the data, the parameter settings and initialization. We then study the influence of some parameters and constraints on the source separation performance. We compare the proposed method with state-of-the-art algorithms in various settings. Finally we evaluate the potential of our method in term of speech alignment.

5.1 Data

We evaluate our approach on synthetic data which consists of three sets: the mixtures, the speech examples, and the training set needed for some baseline approaches. All audio signals are mono and sampled at 16000 Hz.

The *mixture set* consists of eighty different mixtures created as follows. Ten speech signals (five for male voice and five for female voice) in English corresponding to ten different sentences were randomly selected from the test subset of the TIMIT database [26]. Each chosen speech signal was used to produce eight mixtures by adding to it either music or effect background. These background sounds were extracted from real movie audio tracks and the Signal (*speech*) to Noise (*background*) Ratios (SNRs) were set to four different values: -5 dB, 0 dB, 5 dB and 10 dB.

The *example set* is built in accordance to the mixture set. For each of ten TIMIT sentences, twelve corresponding speech examples were created. Two of them were produced via speech synthesizers (one with male voice and one with female voice). [5] Other eight examples were produced by human speakers: two by a female native English speaker, two by a male native English speaker, two by a female non-native speaker (Spanish), and two by a male non-native speaker (French). Each of these speakers produced two examples: the first example by just reading the sentence, and another one by reading the sentence, listening to the mixture, and trying to mimic it. The last two examples were taken from the TIMIT test database, but by different speakers: one male and one female. Note that this example set covers three of the four generating scenarios mentioned in section 2 and schematized on figure 2.

The *training set*, which is used only in several baselines, consists of one hundred spoken sentences from different speakers: fifty males and fifty females. These speech signals were randomly selected from the TIMIT train database.

5.2 Parameter setting and initialization

The STFT is computed using a half-overlapping sine window of length 32 ms (*i.e,* 512 samples). Each column of the $F \times I$ excitation dictionary matrix $\mathbf{W}^e$ is set to a harmonic comb with a given fundamental frequency (pitch). The pitch is varied from 100 Hz to 400 Hz covering mostly frequency range of human speech, with an increment of 1/8 of a tone. The entries in the last column of $\mathbf{W}^e$ are set to the same constant value for representing unvoiced sounds. These settings lead to $I = 186$ columns in $\mathbf{W}^e$. The $F \times L$ matrix $\mathbf{P}$ of spectral blobs, which is used to constrain the dictionary of phonemes and the time-invariant channel filters, is built following the auditory-motivated Equivalent Rectangular Bandwidth (ERB) scale [20]. In our experiment L is set to

[5] We used "ivona" synthesizers `www.ivona.com/en/` to create speech examples.

55. The two matrices $\mathbf{W}^e$ and $\mathbf{P}$ are computed using the Flexible Audio Source Separation Toolbox (FASST) [20] routines.

All parameters in (9) are randomly initialized by positive values, except the synchronization matrix $\mathbf{D}$ that is initialized as described in section 3.4 (see also Fig. 4, bottom-left subplot).

5.3 Influence of some parameters and constraints on source separation performance

We assess the performance of our approach on test data when certain parameters vary and setting various constraints in order to find a good configuration for comparing with state-of-the-art methods. Source separation performance is measured by two well-known metrics: the Signal to Distortion Ratio (SDR) criterion [27] and the Overall Perceptual Score (OPS) criterion [28] computed for the target speech source only.

5.3.1 Study the of influence of trade-off parameter λ and bandwidth B

While the trade-off parameter λ determines the level of guidance by the example, the bandwidth B allows different level of variation in the synchronization between speech example and the mixture signal. We vary these two parameters in our experiment to assess source separation performance as follow:

Source separation performance as a function of bandwidth B. We test our algorithm on the full database mentioned in section 5.1 with different values of B ranging from 1 to 30, and the results are shown in Fig. 5 for both the SDR and the OPS, as well as for low SNRs (-5 dB and 0 dB) and high SNRs (5 dB and 10 dB). We see that for the SDR, there are slight variations (within 0.7 dB) and the maximum of the results averaged over all SNRs is reached for $B = 15$. On the contrary for the OPS, we have a monotone performance increase, and it seems that taking a higher B results in better performance. In our opinion such a behavior can be explained as follows. When the separation is less guided by the example (higher value of B) the reconstructed signal is less constrained and more smooth, which is translated by a better perceptual score. We have fixed $B = 15$ in all following experiments.

Source separation performance as a function of trade-off parameter λ. In this section, in addition to the conventional strategy where the trade-off parameter λ is constant, we also investigate a new strategy where λ is gradually decreasing through the algorithm's iterations starting with some value λ_0 and ending with 0. We hope that such a *constraint relaxation* strategy (the example's influence virtually disappears with $\lambda = 0$) would lead to a better performance, at least for the perceptual score (OPS). We evaluate our algorithm on the full database with different values of constant λ and decreasing λ starting from λ_0. The results are shown in Fig. 6, where the values of λ and λ_0 are also made proportional to N/N' in order to normalize for a possible difference between the mixture length N and the example length N' (for a consistent visualization on Fig. 6 we introduce the parameters $\lambda' = \lambda \times N'/N$ and $\lambda'_0 = \lambda_0 \times N'/N$).

First it should be noted that, in contrast to our expectations, the decreasing λ strategy performs almost the same as the constant λ strategy in terms of both the SDR and the OPS measures, This observation shows that the coupling between the mixture and the speech example is mostly important during the first iterations of the algorithm, i.e., it is not very important whether this constraint is kept or relaxed at the end. This is possibly because the coupling with the speech example is essentially needed to drive a good mixture model initialization. Moreover, we see that the variations of the SDR are quite subtle, but the optimal average performance is reached in two point corresponding to the two different λ variation strategies. In all further experiments we chose the decreasing trade-off parameter strategy with initial $\lambda_0 = 8.5 \times \frac{N}{N'}$. Now if we look at the OPS, we see that the lower the trade-off parameter, the better the performances, and it is possibly because of the same reason than for the bandwidth. Indeed, a lower trade-off parameter corresponds to a smaller amount of guidance by example.

5.3.2 Study the effect of the channel filter

Within the proposed NMPCF-based framework this experiment studies the effectiveness of newly introduced excitation-filter-channel (EFC) speech model compared to the existing excitation-filter (or source-filter) model used in most existing works [18,19]. For that purpose, we compare the source separation performance achieved using the proposed EFC model and that obtained using the excitation-filter model (named EF). The implementation of this EF model within the proposed NMPCF-based approach is simply achieved by fixing channel vectors $\mathbf{w}_Y^c$ and $\mathbf{w}_S^c$ to be vectors of ones (see Eq. (8)) and by skipping their updates (in fact the updates of $\mathbf{e}_Y^c$ and $\mathbf{e}_S^c$) in the MU rules described in section 4.2. The evaluation is done on the full database and the results are summarized in Table 1. We clearly see that the mod-

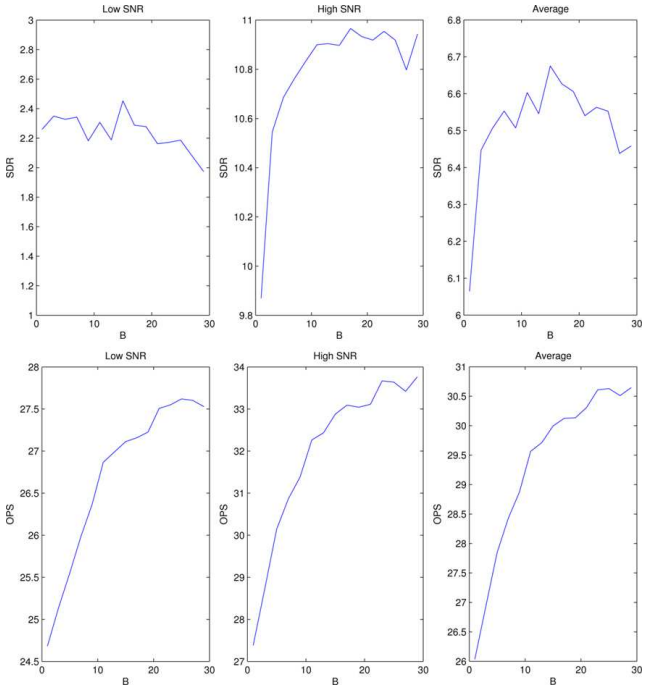

Fig. 5 Performance evolution with respect to bandwidth B. The results are given for two performance measures: the SDR (top) and the OPS (bottom). Left column: the average over low SNRs (-5 dB and 0 dB), middle column: the average over high SNRs (5 dB and 10 dB), and right column: the average over all SNRs.

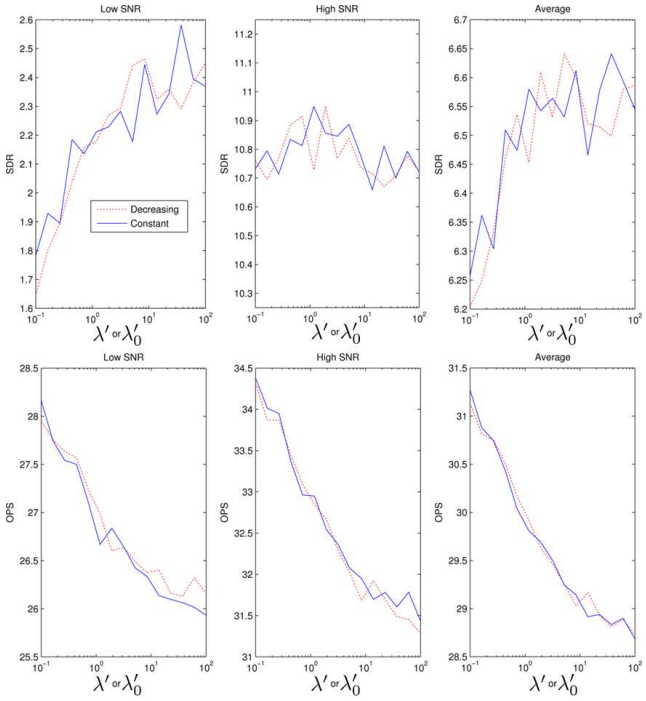

Fig. 6 Performance evolution with respect to constant trade-off parameter λ (solid line) and decreasing trade-off parameter λ starting from λ_0 (dotted line). The results are given for two performance measures: the SDR (top) and the OPS (bottom). Left column: the average over low SNRs (-5 dB and 0 dB), middle column: the average over high SNRs (5 dB and 10 dB), and right column: the average over all SNRs.

eling of channel information greatly improves the overall performance in both low SNR and high SNR conditions, especially in terms of SDR. It is explained by the fact that the channel part of the model captures the time-invariant part of the filter, and whereas in the EF configuration this time-invariant part is shared between the mixture and the model, it is not shared in the EFC configuration where they share only the time-varying part. It is indeed logic to share only the time-varying part between the two models because it really corresponds to the linguistic part of the speech, whereas the time-invariant part corresponds to the recording conditions and intrinsic characteristics of the speaker's vocal tract. The speaker of the example and the one of the mixture are not recorded in the same conditions and they may have very different intrinsic characteristics, thus forcing the time-invariant part of the filter to be the same in the two models would lead to a too strong constraint.

Method	low SNR	high SNR	Avg
EF	0.09 | 26.7	4.58 | 28.7	2.33 | 27.7
EFC	**2.65** | **27.7**	**11.15** | **33.3**	**6.90** | **30.5**

Table 1 Influence of the channel presence on the average performance. Results are shown in the form of SDR | OPS measure and the highest value of each column is in bold.

5.3.3 Study the influence of structural constraints

In this section we investigate structural constraints introduced in section 3.5. The configuration in which the matrices $\mathbf{H}_Y^e$, $\mathbf{H}_S^e$ and $\mathbf{H}_Y^\phi$ are constrained to have only one non-zero entry per column is called GSMM. This setting means that we allow the model of the speech in the mixture and the model of the example to use only one pitch at a time, as well as only one phoneme spectral envelope at a time. The configuration in which we allow only one monotonous path to be non-zero in the matrix $\mathbf{D}$ is called D-Struct. This setting means that the synchronization matrix maps the frames dynamically along the iterations, and it can be viewed as resynchronizing at each iteration. For the D-Struct configuration we initialize $\mathbf{D}$ by doing a DTW and then we allow the path to vary in a band of width $B = 15$ around this initial path. We compared these two configurations with the basic NMPCF-based unconstrained method (the one described in section 3.4) and the results are shown in Table 2. We see that the D-Struct configuration leads to a decrease of performance. On the other hand, the GSMM configuration improves the performance in terms of the SDR, and especially for low SNRs, *i.e.*, when the separation is difficult. This means

that for a more difficult separation problem (low SNR as opposed to high SNR) a more constrained model is beneficial. Finally, in line with our previous observations the best OPS is always achieved by the most unconstrained model, *i.e.*, the basic NMPCF.

Method	low SNR	high SNR	Avg
NMPCF	2.65 | **27.7**	**11.15** | **33.3**	6.90 | **30.5**
"D-Struct"	1.32 | **27.7**	8.10 | 31.0	4.71 | 29.4
GSMM	**4.44** | 25.1	10.36 | 27.9	**7.40** | 26.5

Table 2 Influence of structural constraints on the average performance. Results are shown in the form of SDR | OPS measure and the highest value of each column is in bold.

5.3.4 Multichannel case

We extended our method to the multichannel case by combining the proposed NMPCF-based spectral model with the spatial model described in [21]. This is actually quite straightforward since the proposed single-channel approach acts only in the spectral domain, whereas the one from [21] acts only in the spatial domain. Therefore, in the same spirit as it is done in the general source separation framework from [20], the two can simply be "plugged" together to be used jointly. This extension to the multichannel case was done to participate in the Signal Separation Evaluation Campaign (SiSEC 2013). Due to lack of space the corresponding results as well as the those of other SiSEC 2013 participants are omitted here. However the interested reader can always find them in [22].

5.4 Comparison with different state-of-the-art methods

In this section we compare the performance of the proposed approach with several relevant baselines and state-of-the-art methods.

Baselines non-informed by example The following approaches use neither speech example nor text information[6]:

- NMF: A standard NMF-based method with a general voice spectral dictionary $\mathbf{W}_S \in \mathbb{R}_+^{F\times J}$ ($J = 128$) which is first learned on the training set described in Section 5.1, and then fixed during the parameter estimation.
- EFC-N: A method using the EFC mixture model (5) in a non-supervised manner, as in [19], *i.e.*, filter matrices $\mathbf{W}_S^\phi$ and $\mathbf{H}_S^\phi$ are left free and not coupled with the example. In other words, this method corresponds to the proposed approach with $\lambda = 0$ in (11).
- EFC-S: A method using the EFC mixture model (5), which however is not supervised by example any more, but by our training data. In this approach filter dictionary $\mathbf{W}_S^\phi$ is pre-learned on the training set and then fixed during parameter estimation, while $\mathbf{H}_S^\phi$ is updated.

Baseline informed by example We also consider as a baseline the SbH PLCA-based method [7], within the proposed general workflow, as shown on Fig. 1. Since the mixture $\mathbf{V}_X$ and the example $\mathbf{V}_Y$ are not aligned in general, we used $\mathbf{V}'_Y = \mathbf{V}_Y\mathbf{D}_0$ as example for SbH, where $\mathbf{D}_0$ is the initial synchronization computed with DTW as described in section 3.4. In other words, $\mathbf{D}_0$ is the initial matrix $\mathbf{D}$ with $B = 1$. The SbH itself was implemented following [7]. This baseline is referred hereafter as SbH-DTW.

We compare these baseline methods with different configurations of our method, namely the most simple NMPCF configuration and the GSMM configuration. Table 3 shows average results for different mixture types in terms of both the SDR and the OPS measures.

We can see that the proposed method gives better average results than all the baselines, and especially on the mixtures with low SNRs (difficult cases). These results confirm that textual information is relevant to improve source separation performances. Indeed, it is not surprising that our method performs better than the NMF baseline, which is very basic and does not use a specific model for the speech signal. The fact that our method performs better than the EFC-N baseline indicates that having a speech example is more important for source separation than having a general speech model. Furthermore, the fact that our approach exhibits better performances than the EFC-S baselines is interesting, because it shows that the linguistic information contained in the example matters, and that it is not sufficient to have a general speech model. Finally, our method performs better than the SbH-DTW baseline, and it shows that it is advantageous to model explicitly the time and frequency variations between the example and the speech in the mixture. Moreover, we have a feeling that if the alignment between the mixture and the speech example were improved, the results would be even better.

Finally, an investigation of the performance of the proposed basic NMPCF-based method with respect to

[6] We implemented these approaches with help of the FASST [20].

the example type variations (*e.g.*, same vs. different gender between the target and the example speech, native vs. non-native speaker, human vs. speech synthesizer, etc.) can be found in our preliminary study [1].

5.5 Speech alignment evaluation

This experiment aims at evaluating the temporal alignment between the mixture and the speech example resulted from our algorithm. This alignment is reflected in the final estimate of the synchronization matrix $\mathbf{D}$ and can be decoded using the DTW as in [24]. Let us first define a ground truth and a performance measure for this task. All speech signals from the TIMIT database are supplied with the corresponding phonetic transcriptions that can be used as a ground truth for the alignment evaluation. Therefore, for this evaluation we have used the subsets of our mixture and example sets (see Sec. 5.1) corresponding to the signals from TIMIT. As for the performance measure, we considered $\mathbf{D}$ as a time warping matrix (we decoded a path in it, like in the DTW) and evaluated the proportion of time where $\mathbf{D}$ maps phonetically correctly the two sequences, based on the TIMIT transcriptions. This gave us a number between 0 and 1 that we called *alignment score*. We evaluated this score for different settings of our algorithm (those of the previous subsection), and compared it to the initial DTW (on the MFCCs with euclidean distance). Results are shown in Table 4. One can see that, unfortunately and somehow surprisingly, none of the considered advanced methods provides better alignment result than a very simple DTW used for the initialization. This is possibly due to the fact that our method was designed to perform source separation and not alignment in the first place, and alignment is just a by-product of the method. However, these results are interesting and we believe that improving the estimation of the alignment between the mixture and the speech example would lead to a further improvement of source separation performance of our approach. We leave this potential study for future work.

Method	low SNR	high SNR	Avg
DTW	**0.58**	**0.66**	**0.62**
NMPCF	0.39	0.39	0.39
"D-Struct"	0.48	0.53	0.51
GSMM	0.41	0.45	0.43

Table 4 Alignment score of different configurations. The highest value of each column is in bold.

6 Conclusion

In this paper, we presented an informed speech source separation approach that takes into account the available textual information to guide the separation process via generation of a speech example. We proposed a novel NMPCF modeling framework that allows to efficiently handle different types of variations between the speech example and the speech in the mixture. We also extended the framework to allow incorporation of some novel so-called structural constraints that have quite natural physical motivation. We studied extensively the influence of some key model parameters and constraints on the separation quality. We have found that newly proposed GSMM structural constraints applied to both the excitation and the filter activations improves the results. Moreover, the experimental results over various settings confirm the benefit of the proposed approach over both the non-informed NMF-based baseline methods and a SbH state-of-the-art algorithm [7]. We have also evaluated several configurations of the proposed approach in terms of the alignment accuracy between the target speech and the example speech. We have found that none of the proposed advanced approaches improves the alignment accuracy over a basic DTW applied to the MFCC representations of the example and the mixture. Finally, we extended the method to multichannel mixtures and entered into the Signal Separation Evaluation Campaign (SiSEC 2013) [22].

Future work will consist in exploiting some parameters of speech example NMF decomposition as hyper parameters of a prior distribution to guide the separation process. We also plan investigating within the proposed framework the so-called *soft* co-factorization strategies [29], where the the coupled matrices are not shared exactly, but only approximately. Finally, we will look for more appropriate algorithmic strategies for "D-Struct" constraint decoding (*e.g.*, some strategies based on HMM decoding), and we will look for better alignment approaches to use for the synchronization matrix initialization.

Acknowledgements The authors would like to thank S. Ayalde, F. Lefebvre, A. Newson and N. Sabater for their help in producing the speech examples, as well as the anonymous reviewers for their valuable comments.

References

1. L. Le Magoarou, A. Ozerov, and N.Q.K. Duong. Text-informed audio source separation using nonnegative matrix partial co-factorization. In *Machine Learning for Signal Processing (MLSP), 2013 IEEE International Workshop on*, pages 1–6, 2013.

	Method	Speech + Music				Speech + Effects				Avg
		-5dB	0dB	5dB	10dB	-5dB	0dB	5dB	10dB	
Baselines	NMF	-2.60|20.1	2.22|19.8	7.05|18.7	11.58|20.0	-3.39|26.7	0.66|27.6	4.96|30.4	11.61|33.2	4.01|24.5
	EFC-N	-1.82|24.3	3.07|23.4	8.38|22.9	11.74|22.1	-1.38|24.9	5.15|23.0	**10.74**|23.7	**13.70**|24.4	6.20|23.6
	EFC-S	-2.83|23.1	2.34|22.8	7.19|23.0	11.73|20.8	-3.45|25.1	1.40|26.1	5.95|26.9	10.48|27.5	4.10|24.4
	SbH	-2.21|10.5	3.40|14.6	6.80|17.4	7.97|21.3	1.05|20.6	5.63|27.5	8.21|31.2	9.56|32.0	5.05|21.9
	NMPCF	-0.74|**24.4**	4.27|**28.2**	8.83|**31.6**	**12.29**|**34.0**	0.67|**27.4**	6.40|**30.6**	10.36|**32.7**	13.11|**34.7**	6.90|**30.4**
	GSMM	**2.22**|22.6	**6.52**|25.2	**9.67**|27.2	11.16|28.7	**2.18**|25.6	**6.84**|27.1	9.43|27.6	11.18|28.2	**7.40**|26.5

Table 3 Comparison of different configurations of the proposed method with different baselines. Results are shown in the form of SDR|OPS measures and the highest value of each column is in bold.

2. E. Vincent, S. Araki, F. Theis, G. Nolte, P. Bofill, H. Sawada, A. Ozerov, V. Gowreesunker, D. Lutter, and N. Q. K. Duong. The signal separation evaluation campaign (2007-2010): Achievements and remaining challenges. *Signal Processing*, 92(8):1928–1936, 2012.
3. J. Ganseman, G. J. Mysore, J.S. Abel, and P. Scheunders. Source separation by score synthesis. In *Proc. Int. Computer Music Conference (ICMC)*, pages 462–465, New York, NY, June 2010.
4. R. Hennequin, B. David, and R. Badeau. Score informed audio source separation using a parametric model of non-negative spectrogram. In *Proc. IEEE Int. Conf. on Acoustics, speech, and signal processing (ICASSP)*, pages 45–48, Prague, Czech Republic, 2011.
5. U. Simsekli and A. T. Cemgil. Score guided musical source separation using generalized coupled tensor factorization. In *Proc. 20th European Signal Processing Conference (EUSIPCO)*, pages 2639 – 2643, 2012.
6. J. Fritsch and M. D. Plumbley. Score informed audio source separation using constrained nonnegative matrix factorization and score synthesis. In *Proc. IEEE Int. Conf. on Acoustics, speech, and signal processing (ICASSP)*, pages 888–891, 2013.
7. P. Smaragdis and G. J. Mysore. Separation by "humming": User-guided sound extraction from monophonic mixtures. In *Proceedings IEEE Workshop Applications of Signal Processing to Audio and Acoustics (WASPAA)*, pages 69–72, 2009.
8. D. FitzGerald. User assisted source separation using nonnegative matrix factorisation. In *22nd IET Irish Signals and Systems Conference*, Dublin, 2011.
9. J.L. Durrieu and J.P. Thiran. Musical audio source separation based on user-selected F0 track. In *Proc. Int. Conf. on Latent Variable Analysis and Signal Separation (LVA/ICA)*, pages 438–445, Tel-Aviv, Israel, March 2012.
10. A. Ozerov, C. Févotte, R. Blouet, and J.-L. Durrieu. Multichannel nonnegative tensor factorization with structured constraints for user-guided audio source separation. In *Proc. IEEE Int. Conf. on Acoustics, speech, and signal processing (ICASSP)*, pages 257 – 260, Prague, Czech Republic, May 2011.
11. A. Lefèvre, F. Bach, and C. Févotte. Semi-supervised NMF with time-frequency annotations for single-channel source separation. In *Proc. Int. Symposium on Music Information Retrieval (ISMIR)*, pages 115–120, Porto, Portugal, Oct. 2012.
12. N. J. Bryan and G. J. Mysore. Interactive user-feedback for sound source separation. In *International Conference on Intelligent User Interfaces (IUI)*, Santa Monica, CA, March 2013.
13. Q. K. Duong, Ngoc, Alexey Ozerov, Louis Chevallier, and Joel Sirot. An interactive audio source separation framework based on non-negative matrix factorization. In *Proc. IEEE Int. Conf. on on Acoustics Speech and Signal Processing (ICASSP)*, Florence, Italie, May 2014.
14. S. T. Roweis. One microphone source separation. In *Advances in Neural Information Processing Systems 13*, pages 793–799. MIT Press, 2000.
15. W. Wang, D. Cosker, Y. Hicks, S. Sanei, and J. A. Chambers. Video assisted speech source separation. In *Proc. IEEE Int. Conf. on Acoustics, speech, and signal processing*, pages 425–428, Philadelphia, USA, 2005.
16. G. J. Mysore and P. Smaragdis. A non-negative approach to language informed speech separation. In *Proc. Int. Conf. on Latent Variable Analysis and Signal Separation (LVA / ICA)*, pages 356–363, Tel-Aviv, Israel, March 2012.
17. M. Kim, J. Yoo, K. Kang, and S. Choi. Nonnegative matrix partial co-factorization for spectral and temporal drum source separation. *IEEE Journal of Selected Topics in Signal Processing*, 5(6):1192–1204, 2011.
18. T. Virtanen and A. Klapuri. Analysis of polyphonic audio using source-filter model and non-negative matrix factorization. In *Advances in Models for Acoustic Processing, Neural Information Processing Systems Workshop*, 2006.
19. J. L. Durrieu, G. Richard, B. David, and C. Févotte. Source/filter model for unsupervised main melody extraction from polyphonic audio signals. *IEEE Transactions on Audio, Speech and Language Processing*, 18(3):564–575, 2010.
20. A. Ozerov, E. Vincent, and F. Bimbot. A general flexible framework for the handling of prior information in audio source separation. *IEEE Transactions on Audio, Speech and Signal Processing*, 20(4):1118–1133, 2012.
21. N. Q. K. Duong, E. Vincent, and R. Gribonval. Underdetermined reverberant audio source separation using a full-rank spatial covariance model. *IEEE Transactions on Audio, Speech and Language Processing*, 18(7):1830–1840, Sep. 2010.
22. N. Ono, Z. Koldovsky, S. Miyabe, and N. Ito. The 2013 signal separation evaluation campaign. In *Machine Learning for Signal Processing (MLSP), 2013 IEEE International Workshop on*, pages 1–6, 2013.
23. C. Févotte, N. Bertin, and J.-L. Durrieu. Nonnegative matrix factorization with the Itakura-Saito divergence. With application to music analysis. *Neural Computation*, 21(3):793–830, Mar. 2009.
24. A. Pedone, J. J. Burred, S. Maller, and P. Leveau. Phoneme-level text to audio synchronization on speech signals with background music. In *Proc. INTERSPEECH*, pages 433–436, 2011.
25. D. Ellis. Dynamic time warp (DTW) in Matlab. Web resource, 2003. available: `http://www.ee.columbia.edu/ln/labrosa/matlab/dtw/`.
26. J. Garofolo, L. Lamel, W. Fisher, J. Fiscus, D. Pallett, and N. Dahlgren. DARPA TIMIT: Acoustic-phonetic continuous speech corpus. Technical report, NIST, 1993. distributed with the TIMIT CD-ROM.

27. E. Vincent, R. Gribonval, and C. Fevotte. Performance measurement in blind audio source separation. *IEEE Transactions on Audio, Speech, and Language Processing*, 14(4):1462–1469, July 2006.
28. V. Emiya, E. Vincent, N. Harlander, and V. Hohmann. Subjective and objective quality assessment of audio source separation. *IEEE Transactions on Audio, Speech and Language Processing*, 19(7):2046–2057.
29. N. Seichepine, Slim Essid, Cédric Févotte, and Olivier Cappé. Soft nonnegative matrix co-factorizationwith application to multimodal speaker diarization. In *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3537–3541, 2013.